LEXICAL LINK ANALYSIS APPLICATION: IMPROVING WEB SERVICE TO ACQUISITION VISIBILITY PORTAL

May 14-15, 2014

Dr. Ying Zhao, Dr. Douglas J. MacKinnon, Dr. Shelley P. Gallup,
Research Associate Professors
Distributed Information Systems Experimentation, Naval Postgraduate School

Report Documentation Page		*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **MAY 2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**
4. TITLE AND SUBTITLE **Lexical Link Analysis Application: Improving Web Service to Acquisition Visibility Portal**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Monterey,CA,93943**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES **AFCEA 11th Annual Acquisition Research Symposium, 14-15 May 2014, Monterey, CA.**

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**25**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Background: Critical Needs of Automation, Validation, and Discovery for DoD Acquisition
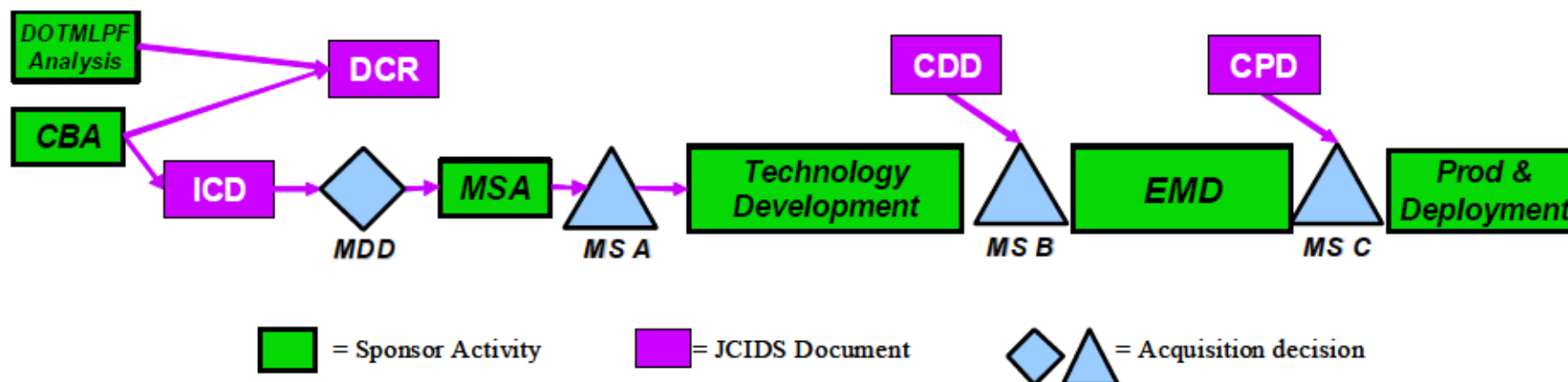
JCIDS Process and Acquisition Decisions
(J-8 CJCSI 3170.01G)(JCIDS, 2009)
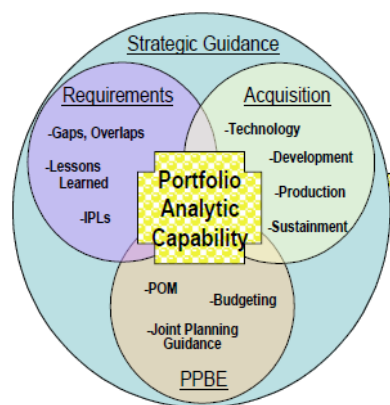
- Data are too voluminous, unformatted, and unstructured!

- Need to leverage automation

  - Extract relations among PE, MDAP, and ACATII
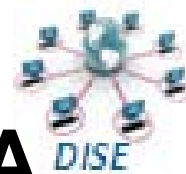
  - Extract costs

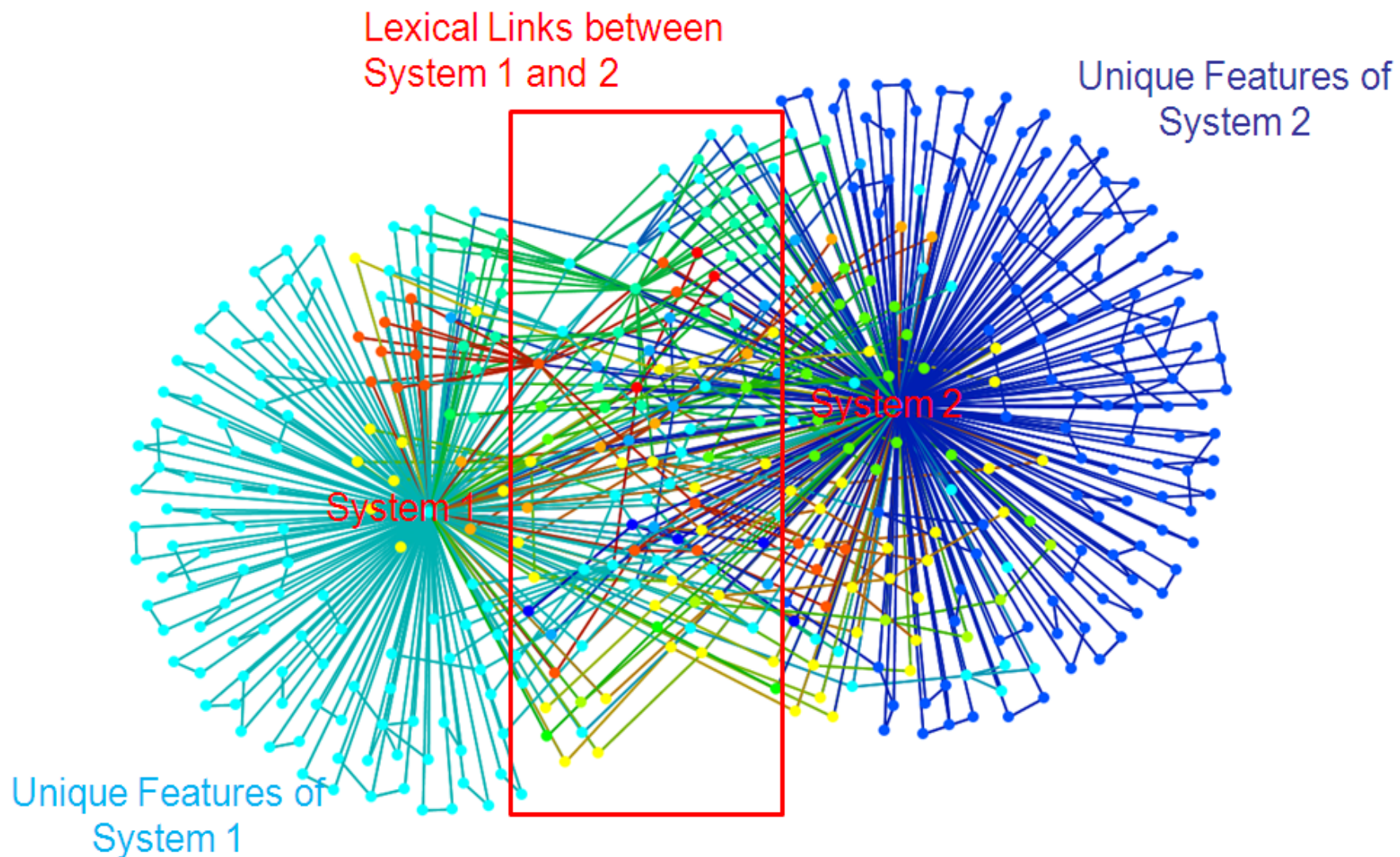# Methods: System Self-awareness (SSA) and Lexical Link Analysis (LLA)

- System Self-awareness
  - The cognitive interface between decision makers and a complex system, expressed in a range of terms or - "features" - a specific vocabulary, or "lexicon," to describe the attributes and surrounding environment of the system.
  - Complex system's ability to assess itself within a global context
    - Authority
    - Expertise
- LLA is a Text Analysis method using bi-gram co-occurrence word pair networks
  - We explored an analytic and visualization of Lexical Link Analysis (LLA), to link warfighter requirements with the acquisition programs and program elements
    - Gallup, MacKinnon, Zhao, Robey & Odell, 2009;
    - Zhao, Gallup & MacKinnon, 2010, 2011a, 2011b, 2011c, 2011d, 2012a, 2012b, 2013;
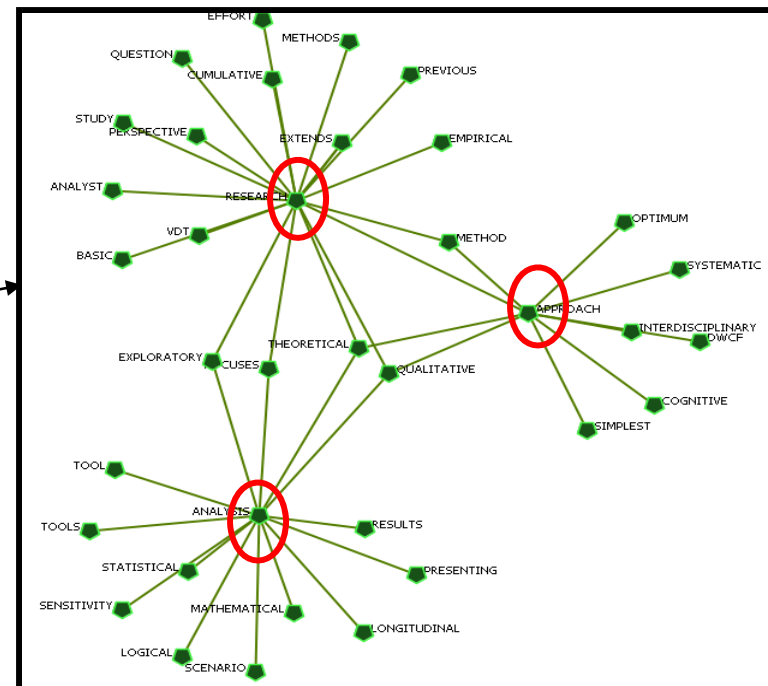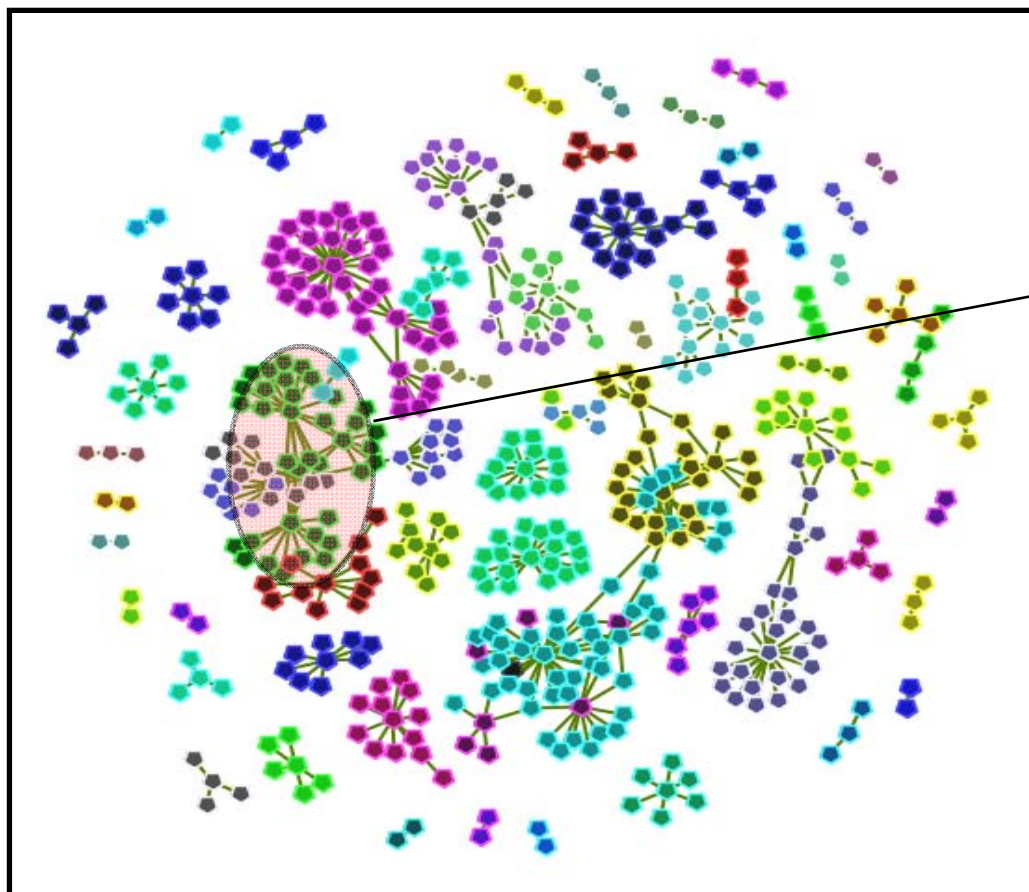   - Zhao, Brutzman & MacKinnon 2013
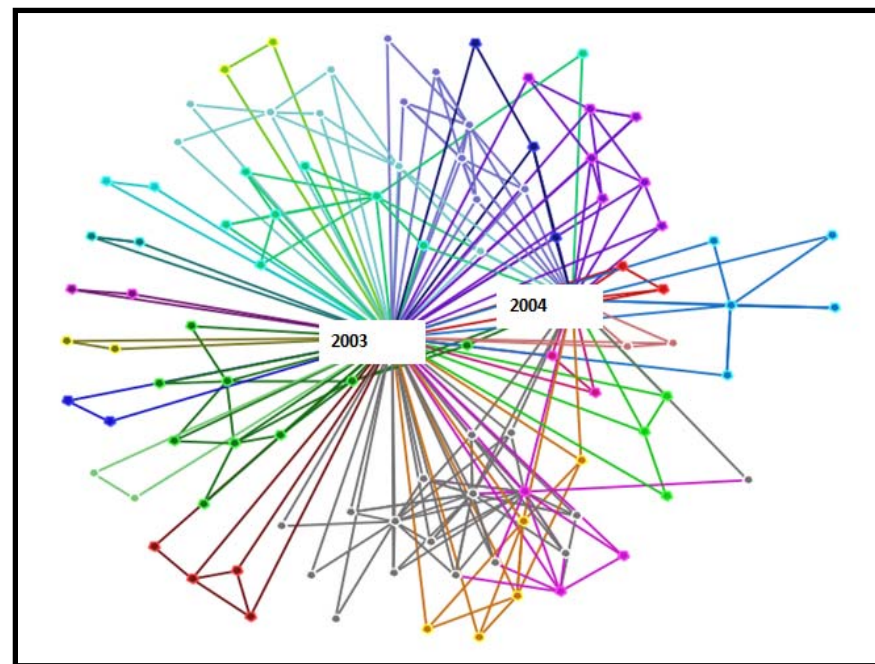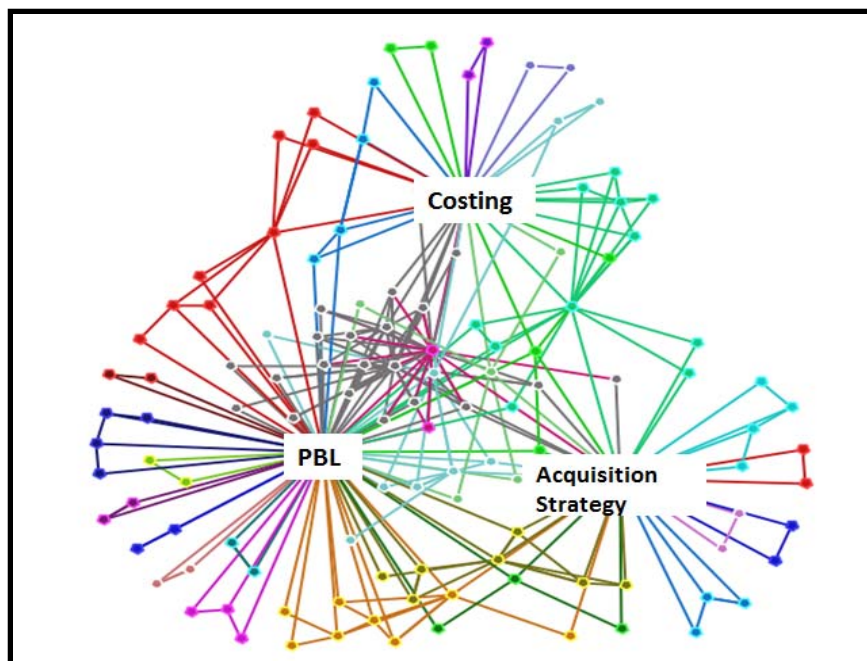
Comparing Two Systems using LLA

Lexical Links between System 1 and 2

Unique Features of System 2

System 1

System 2

Unique Features of System 1

Discovering Themes

Comparing Categories and Time Points

How LLA Methodology Can Help

Warfighters Requirements/Needs (UJTLS)

RDTE Program Elements (DOD Budget $$$ Justification)

• How to validate LLA?

• Do PEs or Programs match requirements?

• Do inter-connected PEs or Programs cost more?

?

LLA automates the possibility to develop awareness of the "fit" between PE's, budget and warfighter requirements.

Weapon Book

(Final Products for Procurement)
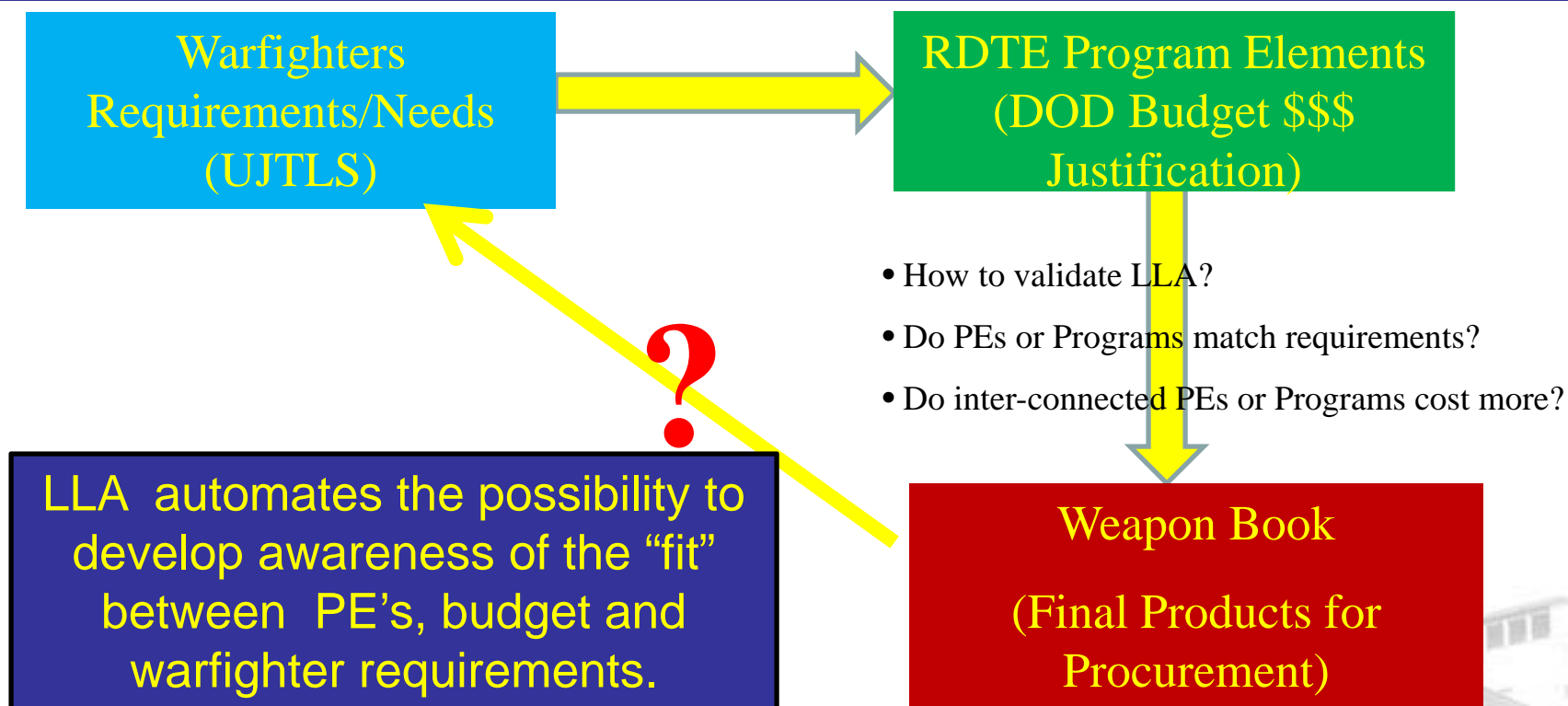
The Acquisition Visibility Portal (AVP)

- DoD acquisition communities and professionals needs to access authoritative and accurate data services for decision-making
  - The Acquisition Visibility Portal (AVP) was such a data service that achieved this purpose by interfacing with Defense Technical Information Center (DTIC)
  - Program Elements: http://www.dtic.mil/descriptivesum/]
  - Warfighter requirements: http:/www.dtic.mil/doctrine]).
  - Defense Acquisition Management Information Retrieval (DAMIR; http://www.acq.osd.mil/damir/)
    - Milestones, costs, schedules, and performance data of selected acquisition reports (SAR)
    - Acquisition Strategy Reports (ASR)
    - …

Gaps and Inconsistencies of AVP Data Sources

- The Office of the Secretary of Defense (OSD) staff review to determine if the program is properly prepared for the next milestone.
  - Thoroughly review these artifacts limited on staffing and little time
  - Each functional community reviews only the particular document for which it is responsible
  - The systems engineering community typically only examines the systems engineering plans (SEP)
  - The test and evaluation community looks only at the Test & Evaluation Master Plans (TEMP)
  - The acquisition community looks at the Acquisition Strategy Reports (ASR).
- Milestone documents
  - Divergent naming conventions may indicate the documents were developed in isolation.
  - Meaningful linkages between these reports, e.g., a capability defined in the acquisition strategy, should be elaborated in the systems engineering plan; the testing of which should be described in the TEMP.
  - Inconsistencies among these documents may reflect a risk to the program.

Where Do the Gaps Come From?
(e.g. Compare ASR and TEMP)

- What are the features or clusters of features (e.g., themes) discussed in ASR but not discussed in TEMP?

- Reasons for gaps

  1. A data quality issue (e.g., a mishandling of data by AVP),

  2. A data classification issue (e.g., unclassified data vs. classified data),

  3. A real requirement gap (i.e., a concept required by acquisition for which no engineering feasibility document or blueprint can be located).

How Might LLA Help?

- Examine large collections of documents for many programs in various categories across the acquisition and engineering communities by

  - Detecting data quality such as inconsistency, gaps, or bad data among categories

  - Identifying data dependencies that might be indicators for program or investment performances and risks

  - Learning from the actual data to see how the common concepts are expressed in different artifacts and communities.

  - Conducting a pair-wise comparison exposes significant disconnections between them.

  - Discovering disconnection or gaps that could be fed back to the human analysts or decision-makers to perform further investigations.

Data Access

- Acquisition Visibility/DMAIR Portal
(https://ebiz.acq.osd.mil/DAMIR/PortalMain/DamirPortal.aspx)
- Acquisition Visibility/AIR Portal
(https://www.dodtechipedia.mil/dodc/plugins/AIR/airdocuments.action)
- Acquipedia https://dap.dau.mil/acquipedia

Automatic Data Streaming

- TEMP: Test & Evaluation Master Plan

- SEP: Systems Engineering Plan

- ASR: Acquisition Strategy Report

- SAR: Selected Acquisition Report

- DAES: Defense Acquisition Executive Summary

- ADM: Milestone B 2366b Certification Acquisition Decision Memorandum

- APB: Acquisition Program Baseline

- TRA: Technology Readiness Assessment

- LCSP: Life Cycle Sustainment Plan

- Acquipedia

- Difficult and follow AVP's proper request processes and rules

LLA Results/Reports: 1. Match Matrix Report

		Match Score	Acquipedia	ASR	LCSP	SEP	DAES	TEMP	ADM	SAR	APB	TRA	Uniqueness Score
1	Acquipedia	943.00	_	499.00(0.15)	521.00(0.22)	292.00(0.19)	86.00(0.05)	82.00(0.06)	25.00(0.04)	51.00(0.09)	3.00(0.01)	5.00(0.01)	3944.00
2	ASR	832.00	499.00(0.15)	_	251.00(0.15)	194.00(0.18)	97.00(0.09)	78.00(0.08)	18.00(0.05)	30.00(0.08)	5.00(0.03)	4.00(0.01)	1415.00
3	LCSP	513.00	521.00(0.22)	251.00(0.15)	_	119.00(0.16)	45.00(0.06)	34.00(0.05)	6.00(0.02)	15.00(0.05)	2.00(0.02)	3.00(0.01)	657.00
4	SEP	239.00	292.00(0.19)	194.00(0.18)	119.00(0.16)	_	34.00(0.07)	11.00(0.02)	6.00(0.03)	10.00(0.06)	0.00(0.00)	3.00(0.02)	253.00
5	DAES	175.00	86.00(0.05)	97.00(0.09)	45.00(0.06)	34.00(0.07)	_	11.00(0.02)	6.00(0.03)	25.00(0.13)	1.00(0.01)	1.00(0.01)	368.00
6	TEMP	86.00	82.00(0.06)	78.00(0.08)	34.00(0.05)	11.00(0.02)	11.00(0.02)	_	0.00(0.00)	2.00(0.01)	0.00(0.00)	0.00(0.00)	353.00
7	ADM	46.00	25.00(0.04)	18.00(0.05)	6.00(0.02)	6.00(0.03)	6.00(0.03)	0.00(0.00)	_	0.00(0.00)	0.00(0.00)	1.00(0.02)	20.00
8	SAR	40.00	51.00(0.09)	30.00(0.08)	15.00(0.05)	10.00(0.06)	25.00(0.13)	2.00(0.01)	0.00(0.00)	_	0.00(0.00)	1.00(0.02)	25.00
9	APB	8.00	3.00(0.01)	5.00(0.03)	2.00(0.02)	0.00(0.00)	1.00(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)	_	0.00(0.00)	7.00
10	TRA	8.00	5.00(0.01)	4.00(0.01)	3.00(0.01)	3.00(0.02)	1.00(0.01)	0.00(0.00)	1.00(0.02)	1.00(0.02)	0.00(0.00)	_	48.00

A match score for a data source is the total number of matched features (e.g., LLA word pairs)

499 is the number of word pairs matched between two sources: Acquipedia and ASR

0.15 is the correlation among these categories between Acquipedia and ASR, normalized using the match score and uniqueness score, computed as $=499/(943+3944)\times 832+1415$

A uniqueness score is the total number of unique word pairs are unique to the source.

2. Correlations Among Data Sources
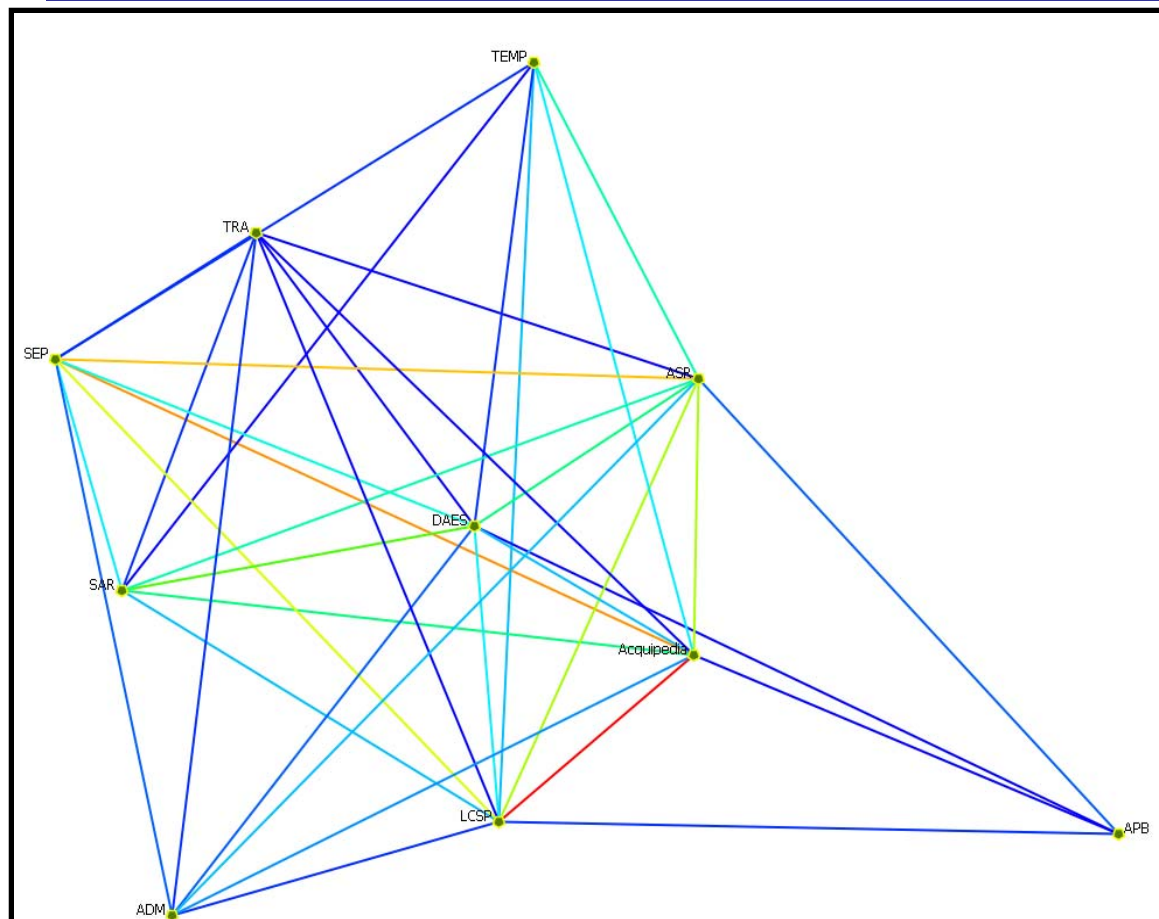
- Ten Data Sources
  - TEMP
  - TRA
  - SEP
  - SAR
  - ADM
  - LCSP
  - DAES
  - ASR
  - APB
  - Acquipedia

- Red – high correlation

- Blue – low correlation

Correlation > 0.1
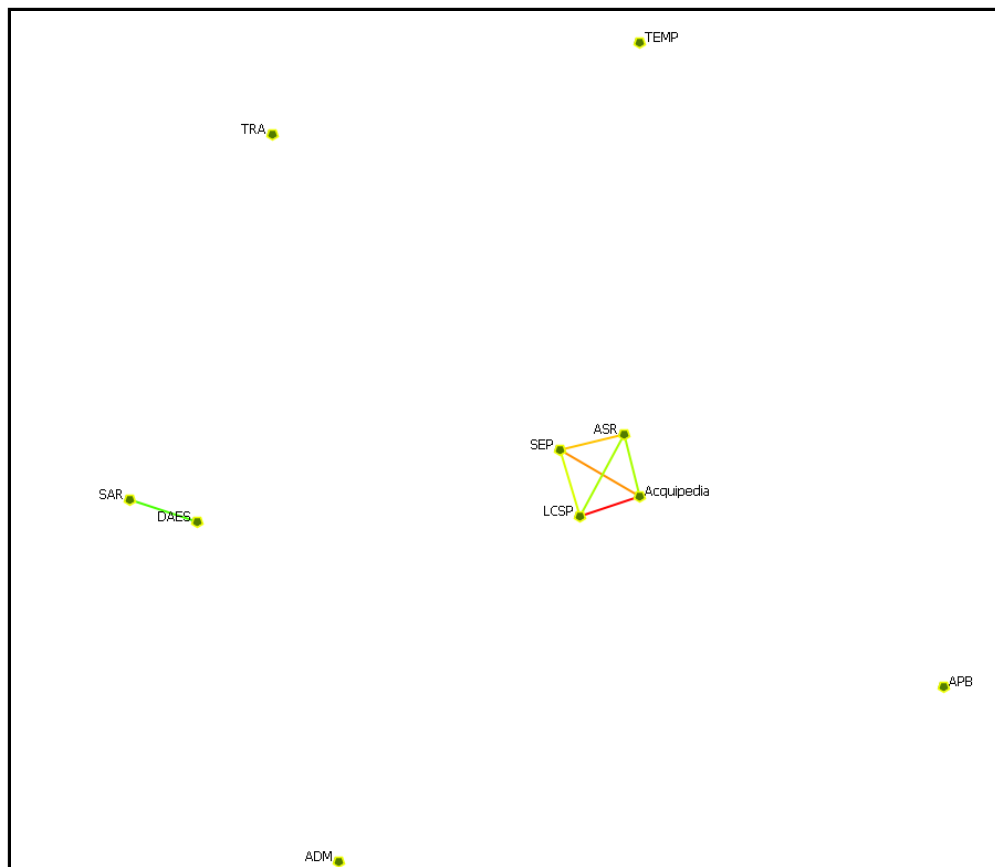
- Data sources LCSP, ASR and SEP have the highest correlations with Acquipedia and with each other.

- ASR, LCSP and SEP may use more standardized vocabularies and terminologies than other data sources.

- SAR and DAES are also correlated with each other more with each other.

3. Drill-down

Frequencies/probability of the word pairs

Click the link to search/drill down to the original documents
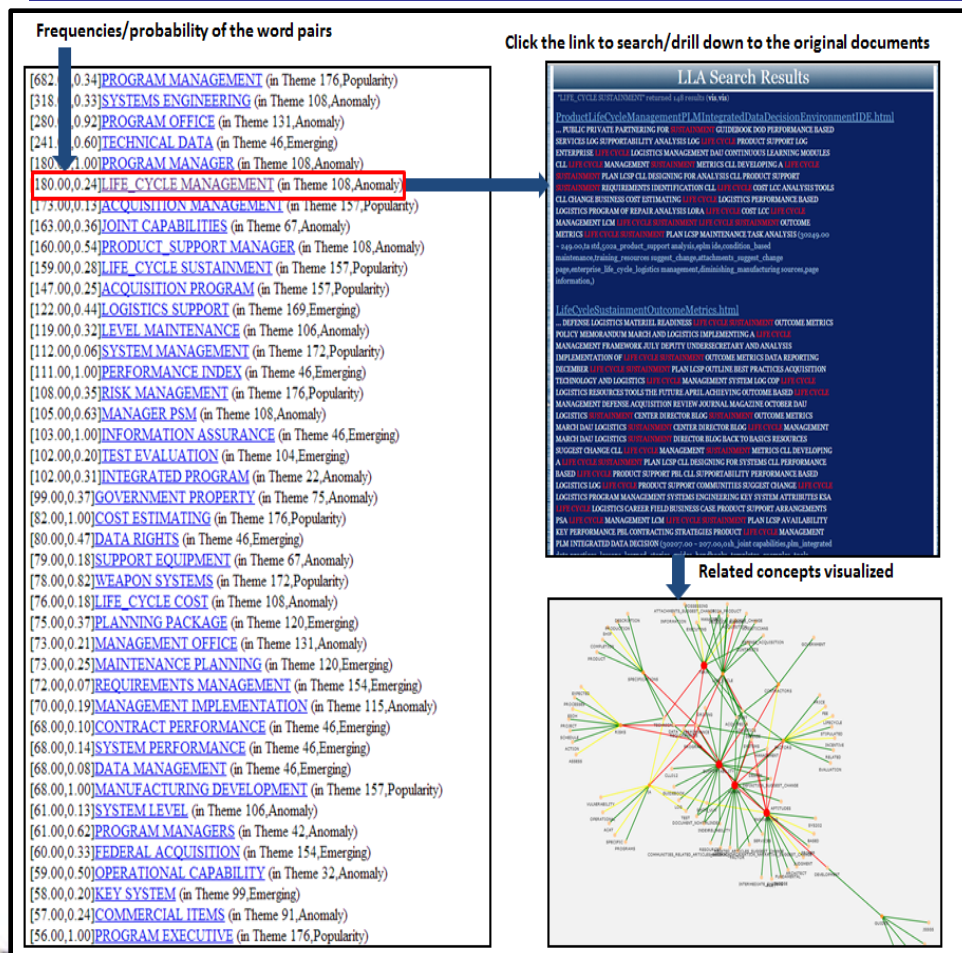
- From Reports of frequencies, bi-gram probabilities, categories of themes

- Allows reach-back search to original documents

- Provides related concepts to become visualized

4. Discovered Theme Report: A List of Themes/Clusters of Word Pairs When Comparing Two Data Sources, e.g., ASR and Acquipedia

| Theme Id	# of Unique Features for Source ASR	# of Unique Features for Source Acquipedia	# of Matched Features	Consensus Rate(Correlation)	Gap Rate	Visualization
157(P)	23	206	34	0.13	0.87	157(P)
176(P)	42	138	47	0.21	0.79	176(P)
167(P)	23	222	16	0.06	0.94	167(P)
172(P)	38	94	23	0.15	0.85	172(P)
169(E)	27	119	31	0.18	0.82	169(E)
46(E)	15	117	46	0.26	0.74	46(E)
104(E)	26	64	12	0.12	0.88	104(E)
120(E)	30	106	27	0.17	0.83	120(E)
111(E)	14	73	8	0.08	0.92	111(E)
154(E)	11	94	21	0.17	0.83	154(E)
99(E)	22	86	27	0.20	0.80	99(E)
141(E)	19	65	7	0.08	0.92	141(E)
179(A)	29	87	27	0.19	0.81	179(A)
124(A)	17	54	10	0.12	0.88	124(A)
106(A)	28	72	16	0.14	0.86	106(A)
67(A)	20	59	23	0.23	0.77	67(A)
108(A)	20	86	22	0.17	0.83	108(A)
32(A)	11	61	6	0.08	0.92	32(A)
149(A)	10	81	11	0.11	0.89	149(A)
22(A)	21	68	8	0.08	0.92	22(A)
166(A)	16	81	10	0.09	0.91	166(A)
131(A)	7	80	6	0.06	0.94	131(A)
115(A)	12	106	3	0.02	0.98	115(A)
66(A)	10	59	4	0.05	0.95	66(A)
68(A)	16	58	16	0.18	0.82	68(A)
145(A)	23	70	7	0.07	0.93	145(A)

- Number of unique features for Source 1 (e.g., ASR)

- Number of unique features for Source 2 (e.g., Acquipedia)

- Matched features for both sources

- Correlation of two sources (or consensus rate), i.e., percentage of the features that are matched

- Gap rate: percentage of the features that are not matched

- These statistics show where the two data sources agree or disagree the most (reflected in the themes)

  - Consensus, e.g., 46(E)

  - Disagreement/gap/inconsistency, e.g., 167(P)

  - Clicking on the Visualization column of 46(E) and 167(P) lead to the visualizations of two areas where consensus and gap took place.

5. One Theme Detail View for Data Source LCSP and ASR
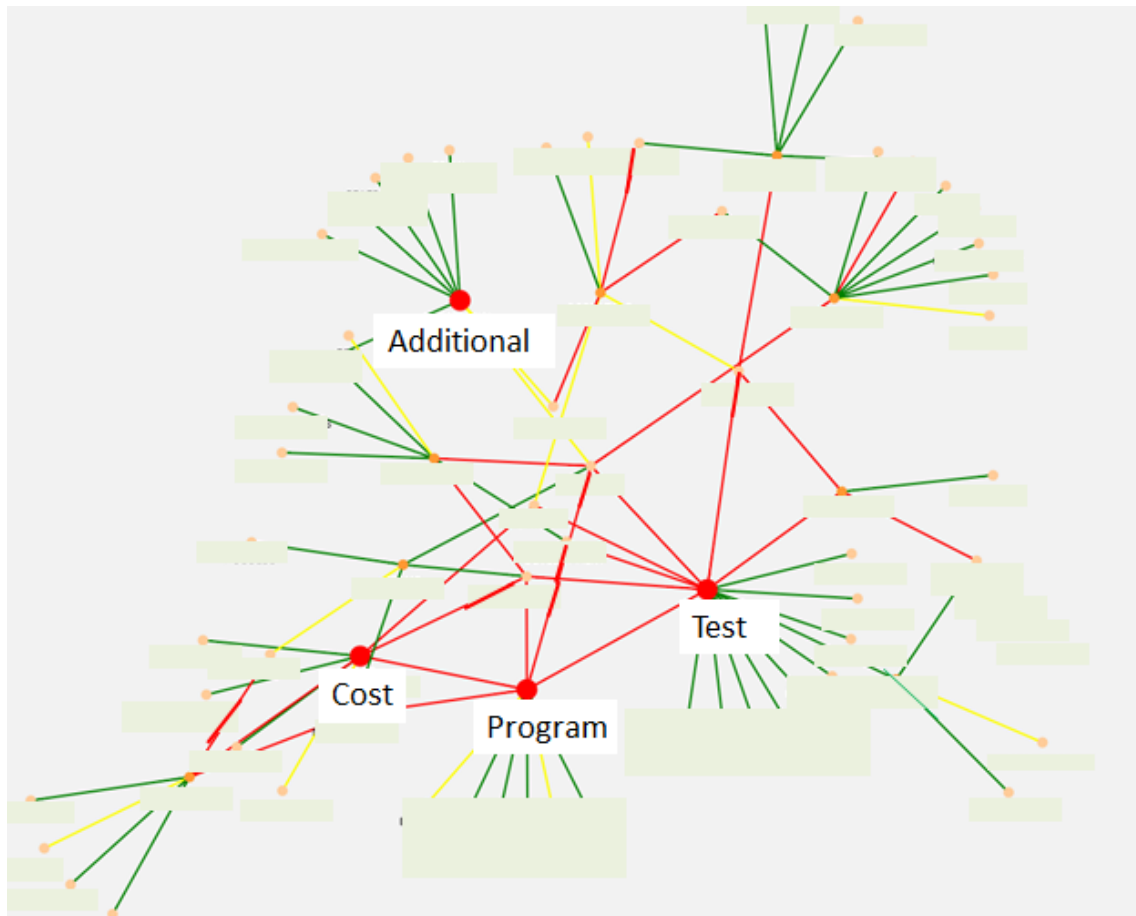
- Red nodes show the most "central" nodes, used as keywords to summarize this theme, i.e. "Additional.Program, Cost, Test,"

- Red Links show word pairs shared by the two sources.

- Yellow Links show the unique word pairs from one source (e.g. LCSP)

- Green links show word pairs from the other source (e.g., ASR).

  o The actual word pairs are eliminated here since the content is FOUO.

- The consensus rate for this theme is 29%, i.e. 29% of word pairs or features are in agreement

- 71% of word pairs are not. As one can see, ASR focuses on "Test" and LCSP does not.

6. Report the List of Word Pairs Matched and Unique in the Two Data Sources (e.g., LCSP and ASR)

- Red nodes: the most "central" nodes, used as keywords to summarize this theme
- Red Links: word pairs shared by at least two sources LCSP,ASR
- Yellow Links: word pairs from Source LCSP
- Green Links: word pairs from Source ASR

Total Consensus Features for LCSP,ASR: 26
Total Uniquness Features for LCSP: 14
Total Uniquness Features for ASR: 49
Gap Rate: 0.71
Consensus Rate: 0.29

Consensus Features for LCSP,ASR	Unique Features for LCSP	Unique Features for ASR

- The actual word pairs are eliminated here since the content is for official use only.

Acquipedia: LLA generated 910 word pairs (out of 5062 total word pairs) matched from this ship-building program's nine data sources. The list can be considered as a set of standardized vocabularies and terminologies

Appendix A: Word Pairs in Acquipedia Matched in a Ship-building Program

[682.00,0.34]PROGRAM MANAGEMENT (in Theme 176,Popularity)
[318.00,0.33]SYSTEMS ENGINEERING (in Theme 108,Anomaly)
[280.00,0.92]PROGRAM OFFICE (in Theme 131,Anomaly)
[241.00,0.60]TECHNICAL DATA (in Theme 46,Emerging)
[180.00,1.00]PROGRAM MANAGER (in Theme 108,Anomaly)
[180.00,0.24]LIFE CYCLE MANAGEMENT (in Theme 108,Anomaly)
[173.00,0.13]ACQUISITION MANAGEMENT (in Theme 157,Popularity)
[163.00,0.36]JOINT CAPABILITIES (in Theme 67,Anomaly)
[160.00,0.54]PRODUCT SUPPORT MANAGER (in Theme 108,Anomaly)
[159.00,0.28]LIFE CYCLE SUSTAINMENT (in Theme 157,Popularity)
[147.00,0.25]ACQUISITION PROGRAM (in Theme 157,Popularity)
[122.00,0.44]LOGISTICS SUPPORT (in Theme 169,Emerging)
[119.00,0.32]LEVEL MAINTENANCE (in Theme 106,Anomaly)
[112.00,0.06]SYSTEM MANAGEMENT (in Theme 172,Popularity)
[111.00,1.00]PERFORMANCE INDEX (in Theme 46,Emerging)
[108.00,0.35]RISK MANAGEMENT (in Theme 176,Popularity)
[105.00,0.63]MANAGER PSM (in Theme 108,Anomaly)
[103.00,1.00]INFORMATION ASSURANCE (in Theme 46,Emerging)
[102.00,0.20]TEST EVALUATION (in Theme 104,Emerging)
[102.00,0.31]INTEGRATED PROGRAM (in Theme 22,Anomaly)
[99.00,0.37]GOVERNMENT PROPERTY (in Theme 75,Anomaly)
[82.00,1.00]COST ESTIMATING (in Theme 176,Popularity)
[80.00,0.47]DATA RIGHTS (in Theme 46,Emerging)
[79.00,0.18]SUPPORT EQUIPMENT (in Theme 67,Anomaly)
[78.00,0.82]WEAPON SYSTEMS (in Theme 172,Popularity)
[76.00,0.18]LIFE CYCLE COST (in Theme 108,Anomaly)
[75.00,0.37]PLANNING PACKAGE (in Theme 120,Emerging)
[73.00,0.21]MANAGEMENT OFFICE (in Theme 131,Anomaly)
[73.00,0.25]MAINTENANCE PLANNING (in Theme 120,Emerging)
[72.00,0.07]REQUIREMENTS MANAGEMENT (in Theme 154,Emerging)
[70.00,0.19]MANAGEMENT IMPLEMENTATION (in Theme 115,Anomaly)
[68.00,0.10]CONTRACT PERFORMANCE (in Theme 46,Emerging)
[68.00,0.14]SYSTEM PERFORMANCE (in Theme 46,Emerging)
[68.00,0.08]DATA MANAGEMENT (in Theme 46,Emerging)
[68.00,1.00]MANUFACTURING DEVELOPMENT (in Theme 157,Popularity)
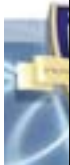
Distribution of Acquipedia Features Among the 910 Matched Features

	Percentage	Number of Features
Anomalous features	39%	358
Emerging features	35%	322
Popular features	25%	230

This validates that our observations : Anomalous and emerging features are more interesting because they are used in the documents regarding an actual ship-building program.

- Popular or Normal (P):

  - Themes contain the highest number of mutually connected word pairs.

  - Themes represent the main topics in a corpus at the time.

  - May be also regarded as *less interesting* because they are already in the public consensus and awareness, therefore, less room for growth.

- Emerging (E): themes containing the medium number of mutually connected word pairs, these themes may grow to popular over time as we show later in the examples.

- Anomalous (A): themes containing the lowest number of mutually connected word pairs. These themes may be off-topics which may seem they do not belong here compared to other ones and may be interesting for further investigation.

Summary

- LLA was used to analyze an MDAP program using AVP data in which we

  – Demonstrated a set of comprehensive LLA analysis reports and visualizations generated automatically using multiple categories of program data as data sources.

  – Revealed correlations and gaps among at least eight data sources.

  – Formed the basis for further inquiry or future reconciliation of the expectations (e.g., acquisition strategy) and realities (e.g., engineering feasibility) from various communities for the same MDAP program.

  – Discovered in detail where the gaps and inconsistencies of the data across multiple data sources reside which lead to the identification of future specific and productive directions for further examination

Planned Future Work

- Continuing work with Sponsors and AVP analysts to develop a process to generate the LLA reports and visualizations for any given program in AVP.

- Studying the program interactions for a portfolio of programs.

  – Select a portfolio of programs and focuses on one type of data sources, for example ASR to see how LLA can depict the interaction risks.

- Conducting *supervised learning data* to train LLA using Acquipedia to improve the understanding of context-dependent meaning.

Acknowledgements

- We thank Mr. Robert Flowe from OUSD(AT&L)/ARA who provided sponsorship to access the Acquisition Visibility Portal, and relevant questions along with insightful guidance and discussions